Analyse des patients COVID-19 : identification des patients à haut risque

1. Sélection des patients COVID-19

Pour se concentrer sur les complications graves liées à la COVID-19, nous avons **réduit le jeu de données initial** aux patients ayant un test positif au COVID-19. Cela a permis de filtrer les patients non infectés ou avec des tests non concluants, afin d'éviter de biaiser l'analyse.

Dans le code, cette étape est réalisée par :

```
df["COVID_POSITIVE"] = df["CLASIFFICATION_FINAL"].apply(lambda x: 1 if x in [1,2,3] else 2)
df = df[df["COVID_POSITIVE"] == 1]
```

2. Prétraitement et gestion des valeurs manquantes

Le jeu de données original contenait de nombreuses valeurs manquantes, codées 97, 98 ou 99, ainsi que des valeurs manquantes classiques (`NaN`).

- Pour **l'âge**, 97, 98 et 99 peuvent représenter des valeurs réelles (patients très âgés), nous n'avons donc **pas remplacé ces valeurs**.
- Pour les autres variables médicales et binaires, la proportion de valeurs manquantes était faible. Nous avons alors utilisé des méthodes simples :
    - **Mode (valeur la plus fréquente)** pour les variables binaires : diabète, hypertension, obésité, etc.
    - **Médiane** pour l'âge et les variables numériques continues.

Pour INTUBED et ICU, l'imputation a été réalisée par **groupes de patients ayant des caractéristiques similaires** (âge, sexe, comorbidités).
Exemple pour INTUBED :

```
df["INTUBED"] = df.groupby(intubed_var)["INTUBED"].transform(
    lambda x: x.fillna(x.mode()[0] if not x.mode().empty else 2)
)
```

Pour la variable **PREGNANT**, tous les hommes ont été automatiquement codés comme non enceintes (`2`), car la grossesse n'est pas applicable. Cela montre l'importance d'une **imputation médicalement cohérente**.

3. Définition de la variable HAUT_RISQUE

Nous avons défini une variable binaire `HAUT_RISQUE` pour identifier les patients présentant une **issue grave** :

- INTUBED = 1 (intubation)
- ICU = 1 (soins intensifs)
- DEATH = 1 (décès)

La logique est :

```
df["HAUT_RISQUE"] = np.where(
    (df["INTUBED"] == 1) | (df["DEATH"] == 1) | (df["ICU"] == 1),
    1,
    2
)
```

Justification médicale

- L'intubation indique une insuffisance respiratoire sévère.
- L'admission en ICU reflète un état critique nécessitant un suivi intensif.
- Le décès peut survenir même sans ICU ou intubation, par exemple chez des patients fragiles ou avec limitation thérapeutique.

Ainsi, **HAUT_RISQUE = 1 regroupe tous les patients présentant une complication grave**, permettant d'identifier ceux nécessitant une attention particulière.

4. Analyse statistique : crosstab et test de Chi2

Pour chaque variable, nous avons calculé des **tables de contingence (crosstab)** pour voir la répartition des patients ICU et INTUBED selon chaque caractéristique.

Exemple : ICU vs PNEUMONIA

```
PNEUMONIA    1.0    2.0
ICU
1.0        87.06  12.94
2.0        64.43  35.57
```

Puis nous avons utilisé le **test du Chi2** pour évaluer la **significativité de l'association** :

- **Hypothèse nulle (H0)** : pas d'association entre la variable et l'événement (ICU ou INTUBED).
- **Hypothèse alternative (H1)** : il existe une association.

Une **p-value < 0.05** indique une association statistiquement significative.

Résultats clés pour ICU :

- Significatif : PNEUMONIA, INTUBED, AGE, SEX, OBESITY, RENAL_CHRONIC, DEATH
- Non significatif : COPD, ASTHMA, PATIENT_TYPE, TOBACCO

Résultats clés pour INTUBED :

- Significatif : PNEUMONIA, AGE, SEX, DIABETES, OBESITY, RENAL_CHRONIC, ICU, DEATH
- Non significatif : ASTHMA, PATIENT_TYPE

Ces analyses confirment **les facteurs de risque connus cliniquement** pour les complications graves liées à la COVID-19.

5. Modélisation prédictive

Nous avons testé plusieurs modèles supervisés pour prédire HAUT_RISQUE :

- **Logistic Regression** : modélisation linéaire de la probabilité d'être à haut risque.
- **Random Forest** : ensemble d'arbres de décision, robuste aux interactions et non-linéarités.
- **Naive Bayes** : basé sur la probabilité conditionnelle, rapide mais suppose l'indépendance des variables.
- **SVM Linear** : sépare les classes par un hyperplan linéaire, adapté pour des datasets équilibrés.

Évaluation des modèles

Nous avons utilisé les métriques suivantes :

- **Accuracy (précision globale)** : proportion de prédictions correctes
- **Precision** : proportion de patients prévus à haut risque qui le sont réellement
- **Recall (sensibilité)** : proportion de patients à haut risque correctement identifiés

Modèle	Accuracy	Precision	Recall
Logistic Regression	0.6899	0.7141	0.6899
Random Forest	0.6845	0.7135	0.6845
Naive Bayes	0.6525	0.6950	0.6525
SVM Linear	0.6905	0.7128	0.6905

Choix final : Logistic Regression ou SVM Linear

- Légère supériorité en précision et rappel pour SVM Linear, mais la **Logistic Regression** reste très interprétable médicalement (coefficients = poids des facteurs de risque).
- Les deux modèles sont adaptés à la prédiction des patients à haut risque.

6. Conclusion

- La combinaison de l'analyse **statistique (Chi2)** et de l'**imputation médicalement cohérente** permet d'identifier correctement les facteurs associés aux complications graves.
- La variable HAUT_RISQUE intègre **intubation, ICU et décès**, ce qui reflète la réalité clinique : un décès peut survenir sans passage en ICU ou intubation.
- L'utilisation de la **médiane ou du mode pour l'imputation des variables restantes** permet de conserver un jeu de données complet, sans introduire de biais majeur.
- Le modèle de prédiction choisi (Logistic Regression ou SVM Linear) permet d'identifier les patients à risque avec une précision et un rappel proches de 70 %, suffisant pour soutenir la prise de décision clinique et la gestion des ressources hospitalières.